BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

VOLUME I GENERAL PRINCIPLES

MOLECULAR BIOLOGY OF THE GENE

FOURTH EDITION

James D. Watson COLD SPRING HARBOR LABORATORY

Nancy H. Hopkins MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Jeffrey W. Roberts CORNELL UNIVERSITY

Joan Argetsinger Steitz YALE UNIVERSITY

Alan M. Weiner YALE UNIVERSITY

The Benjamin/Cummings Publishing Company, Inc.

Menlo Park, California · Reading, Massachusetts · Don Mills, Ontario
Wokingham, U.K. · Amsterdam · Sydney · Singapore
Tokyo · Madrid · Bogota · Santiago · San Juan

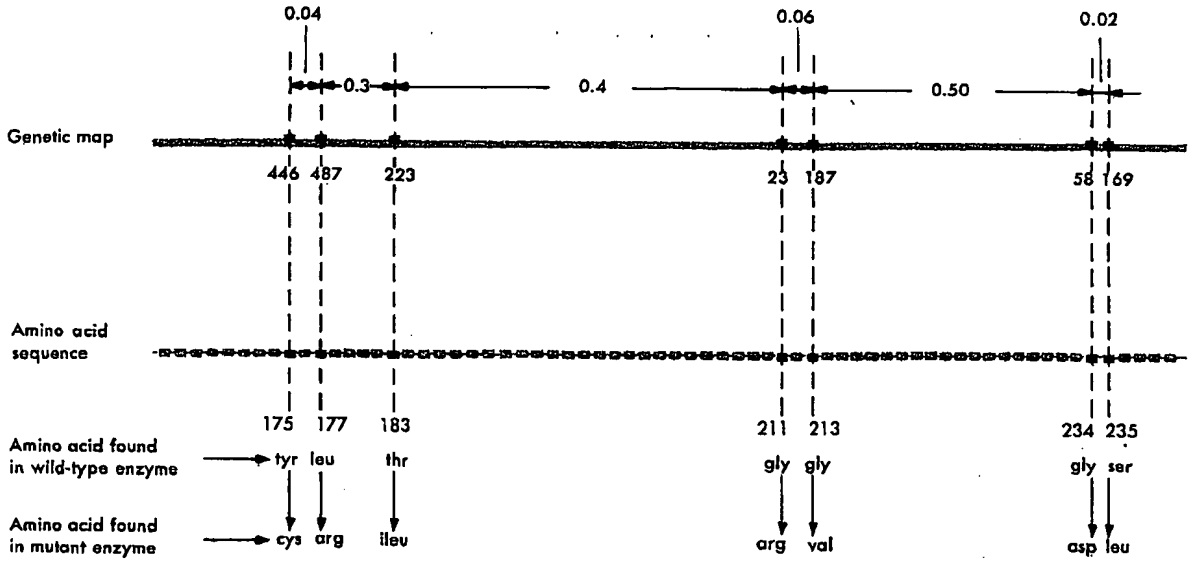
Cover art is a computer-generated image of DNA interacting with the Cro repressor protein of bacteriophage λ. The image was prepared by the Graphic Systems Research Group at the IBM U.K. Scientific Centre.

Editor: Jane Reece Gillen
Production Supervisor: Karen K. Gulliver
Editorial Production Supervisor: Betsy Dilernia
Cover and Interior Designer: Gary A. Head
Contributing Designers: Detta Penna, Michael Rogondino
Copy Editor: Janet Greenblatt
Art Coordinator: Pat Waldo
Art Director and Principal Artist: Georg Klatt
Contributing Artists: Joan Carol, Cyndie Clark-Huegel, Barbara Cousins, Cecile Duray-Bito, Jack Tandy, Carol Verbeek, John and Judy Waller

Library of Congress Cataloging-in-Publication Data
Molecular biology of the gene.
Rev. ed. of: Molecular biology of the gene / James D. Watson. 3rd ed. c1976.
Bibliography
Includes index.
Contents: v. 1. General principles.
1. Molecular biology. 2. Molecular genetics.
I. Watson, James D., 1928– . [DNLM: 1. Cytogenetics. 2. Molecular Biology. QH 506 M7191]
QH506.M6627 1987 574.87'328 86-24500
ISBN 0-8053-9612-8

ABCDEFGHIJ-MU-89876

The Benjamin/Cummings Publishing Company, Inc.
2727 Sand Hill Road
Menlo Park, California 94025



amino acids. This sequence allows us to see how the location of a mutation within a gene is correlated with the location of the replaced amino acid in its polypeptide chain product. Since both genes and polypeptide chains are linear, the simplest hypothesis is that amino acid replacements are in the same relative order as the mutationally altered sites in the corresponding mutant genes. This was most pleasingly demonstrated in 1964. The location of each specific amino acid replacement is exactly correlated with its location along the genetic map, a property called **colinearity.** Thus, successive amino acids in a polypeptide chain are controlled, or coded, by successive regions of a gene.

Figure 8-11
Colinearity of the gene and its protein product: Here is the genetic map for one-fourth of the gene coding for the amino acid sequences in the *E. coli* protein tryptophan synthetase A. The designation 0.04, for example, refers to map distances (frequencies of recombination) between tryptophan synthetase mutations A446 and A487. The numbers in the amino acid sequence refer to their position in the 267 residues of the A protein. Following convention, the amino terminal end of the segment is on the left.

Mutable Sites Are the Base Pairs Along the Double Helix

In all bacterial genes extensively mapped, the large number of linearly arranged mutable sites that have been found in each gene, and between which genetic recombination (crossing over) is possible, leaves us no choice but to conclude that these sites are the specific base pairs along the DNA of the respective gene (Figure 8-12). A given mutable site can thus exist in any of four different states, AT, TA, GC, or CG. Many mutations are therefore likely to represent simple switches from one state to another. The genetic data that reveal deletions and insertions of genetic material must now be thought of in terms of the addition or deletion of discrete blocks of one to very many base pairs. The three classes of mutations resulting from changes in the sequence of nucleotide bases are illustrated in Figure 8-13.

By carefully studying the fine details of genetic maps, we should be able to obtain important information about the corresponding DNA. However, not every change in base sequence leads to easily observed changes in the corresponding protein. In the genetic code, many amino acids are specified by more than one **codon** (set of three adja-

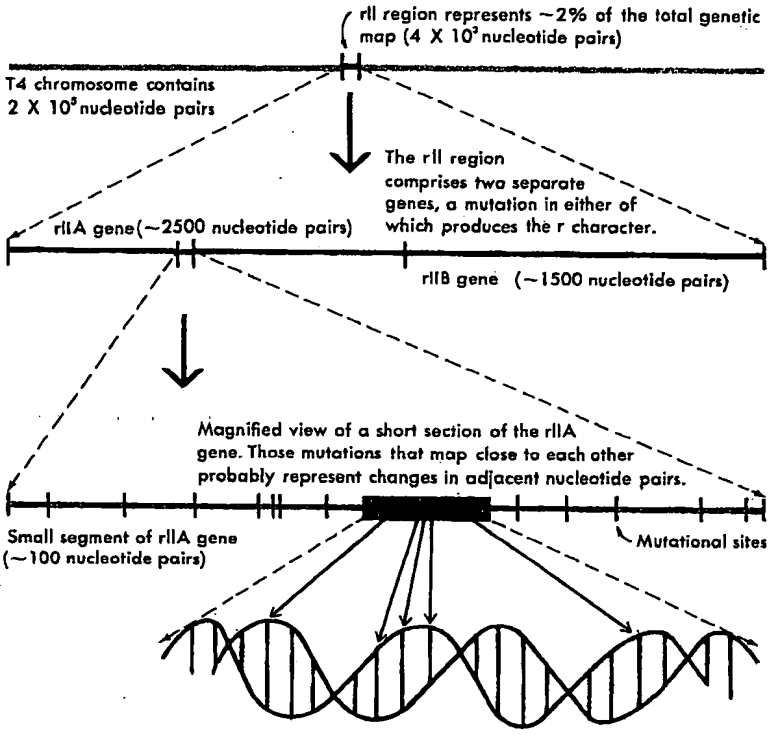


Figure 8-12
The relationship of mutations in the *rII* region of the phage T4 chromosome to the structure of DNA.

cent bases), which means that in many cases, base-pair substitutions will not lead to any amino acid replacements. Moreover, as we document later, many of the amino acids in proteins are not essential, and when they are replaced by somewhat similar amino acids, the proteins often retain full activity. The number of observed mutable sites therefore seriously underrepresents the number of base pairs within the corresponding gene.

There Are Four Alternative Structures for Each Mutable Site[8,9]

As anticipated, enzymatically inactive tryptophan synthetase molecules resulting from independent mutations at the same mutable site (as shown by failure to give wild-type recombinants) do not always contain the same amino acid replacement. For example, changes in a single mutable site that specifies the amino acid at position 213 results in the replacement of glycine by either glutamic acid or valine. Inspection of the genetic code (see Chapter 15) indicates that in the wild-type strain, this glycine must be specified by either GGA or GGG codons and that the mutable site under study specifies the G in the middle position of this codon. When this G is replaced by U, valine (GUA or GUG) becomes inserted into the glycine site while its replacement by A generates the glutamic acid (GAA or GAG) substitution. Further study of this particular mutable site might eventually turn up the anticipated third replacement in which a G to C switch leads to the appearance of alanine (GCA or GCG).

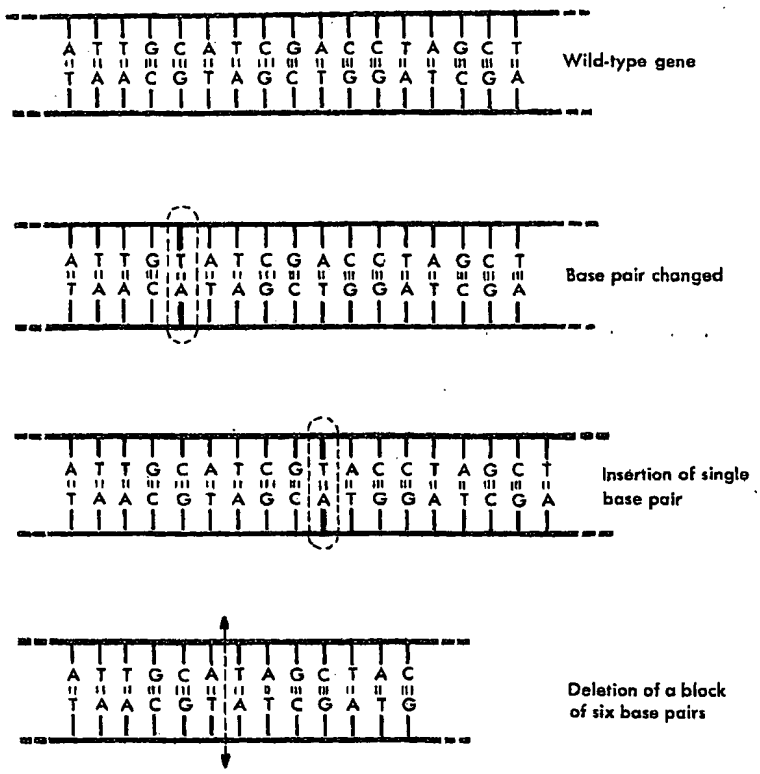


Figure 8-13
Three classes of mutations result from introducing defects in the sequence of bases (A, T, G, C) attached to the backbone of the DNA molecule. In one class, a base pair is simply changed from one into another (i.e., GC to AT). In the second class, a base pair is inserted (or deleted). In the third class, a block of base pairs is deleted (or inserted).

Single Amino Acids Are Specified by Several Adjacent Nucleotide Bases

We expected to find that given amino acids within a particular protein are specified by adjacent mutable sites. This point was first demonstrated in the tryptophan synthetase A gene, where the relevant evidence came from study of the tryptophan synthetase fragment illustrated in Figure 8-14. Treatment of the wild-type strain with a mutagen had given rise to mutant A23, in which arginine replaces glycine (this time at position 212), and mutant A46, in which glutamic acid replaces glycine at the same position. The difference between A23 and A46 does not represent changes to alternative forms of the same mutable site, since a genetic cross between A23 and A46 yields a number of wild-type recombinants (glycine in position 212). If these changes were at the same mutable site, no wild-type recombinants would be produced. Moreover, the very low observed frequency of the wild-type recombinants is compatible with the prediction from the genetic code that these mutable sites are adjacent to each other.

Additional genetic evidence that confirms the separate locations of the A23 and A46 mutable sites comes from observing how A23 and A46 themselves mutate upon treatment with mutagens. After exposure to a mutagen, both strains give rise to new strains, some of which contain active tryptophan synthetase A chains with glycine in position 212. These reverse mutations most likely involve changing the altered mutable sites back to the original wild-type configuration. However, strains containing active tryptophan synthetase also arise

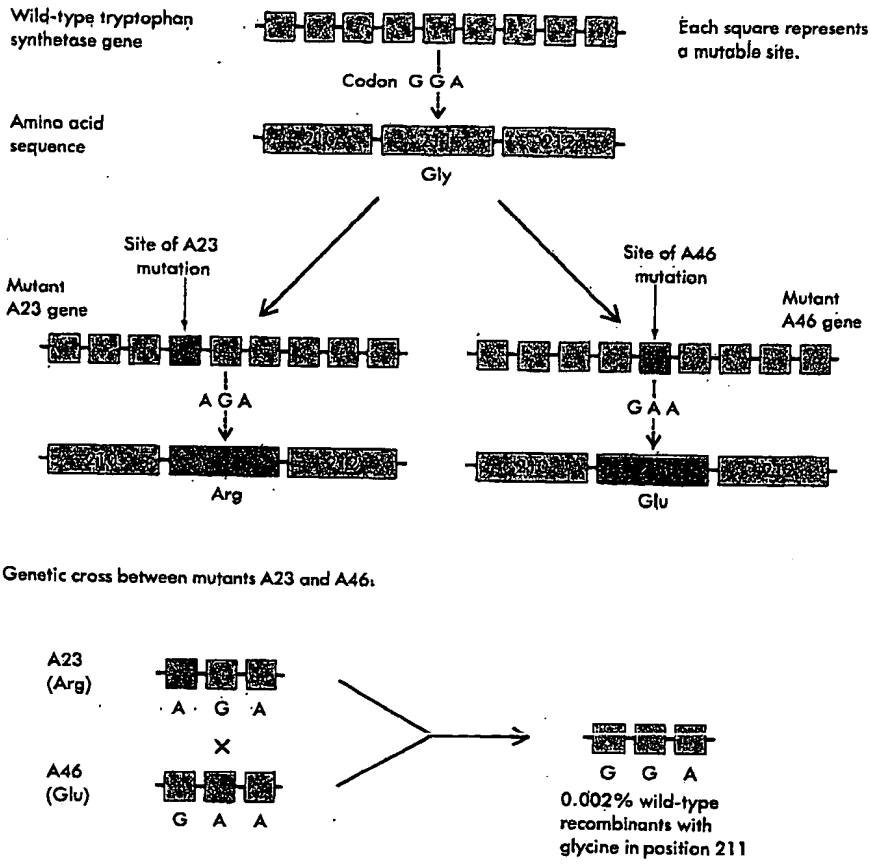


Figure 8-14
Demonstration that a single amino acid is specified by more than one mutable site. We now know that the mutable sites are DNA bases and the codons are actually bases complementary to these in mRNA. (After Emanual J. Murgola.)

in which the amino acid in position 212 is replaced by another amino acid. Most significantly, the type of replacement differs for strains A23 and A46. Besides back-mutating to glycine, strain A23 mutates to threonine and serine, whereas A46 mutates to alanine and valine in addition to glycine. The failure of A23 ever to give rise to alanine or valine and the failure of A46 ever to mutate to threonine or serine is very difficult to explain if their differences from wild type are based on alternative configurations of the same mutable site. But these mutational patterns make perfect sense if glycine at the 212 position is coded by GGA with the A23 mutation to arginine representing a G to A change at the first position of the codon to give rise to AGA and the A46 mutation to glutamic acid occurring at the middle (second) position to give rise to GAA. Their divergent subsequent mutations to serine and threonine and to alanine and valine, respectively, can also be understood by inspecting the genetic code (Figure 8-15).

Single Amino Acid Substitutions Usually Do Not Alter Enzyme Activity

The ability of a polypeptide chain to be enzymatically active does not require an exactly specified amino acid sequence. This is shown by examination of the new mutant strains obtained by treating strains A23 and A46 with mutagens. The possession of either glycine or serine in position 212 yields a fully active enzyme, whereas threonine in

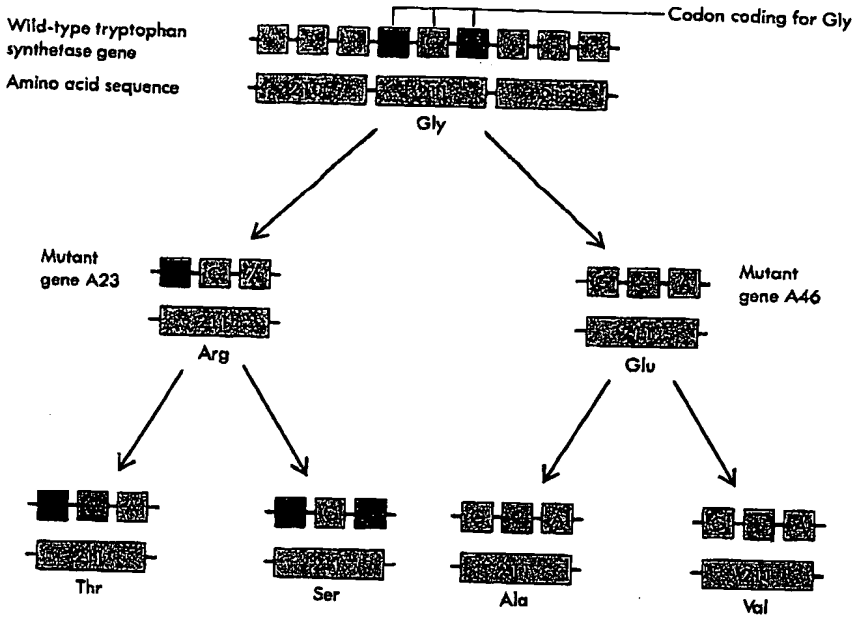


Figure 8-15
Formation of mutants A23 and A46 and their subsequent mutations. Notice that Thr and Ser cannot result from a single base change to the codon for Glu; likewise, Ala and Val cannot result from only one base change to the codon for Arg. Therefore, the A23 and A46 mutants must occur from mutations at two different mutable sites, as shown in Figure 8-14.

the same position yields an enzyme with reduced activity, demonstrating that the activity of an enzyme does not demand a perfectly unique amino acid sequence (Figure 8-16). In fact, evidence now indicates that amino acid replacements in many parts of a polypeptide chain can occur without seriously modifying catalytic activity. However, one sequence may often be best suited to a cell's particular needs, and it is this sequence that is encoded by the wild-type allele. Even though other sequences are almost as good, they will tend to be selected against in evolution.

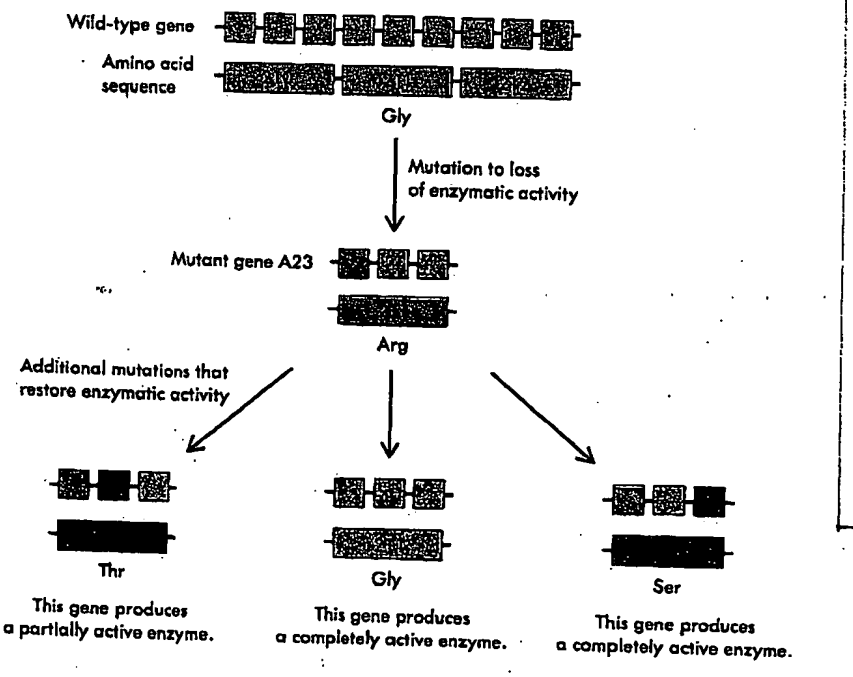


Figure 8-16
Evidence that many amino acid replacements do not result in loss of enzymatic activity.

A Second Amino Acid Replacement May Cancel Out the Effect of the First[10]

The conclusion that minor changes to amino acid sequence do not significantly alter enzyme activity is extended by the finding that some mutations that convert inactive mutant enzymes to active forms may work by causing a second amino acid replacement in the mutant enzyme. Consider mutant A46, which produces inactive tryptophan synthetase because of the substitution of glutamic acid for glycine at protein 212. In this case, distant second-site mutations that result in the active enzyme occasionally emerge. For example, the second-site mutation A446 is located one-tenth of a gene length away from the first mutation. The double mutant A46A446 produces active enzyme molecules containing two amino acid replacements: the original glycine-to-glutamic acid shift and a tyrosine-to-cysteine shift located 36 amino acids away (Figure 8-17).

The second shift can be studied independently of the first by obtaining recombinant cells with only the A446 mutation. Most interestingly the A446 change, when present alone, also results in an inactive enzyme. We thus see that a combination of two wrong amino acids can produce an enzyme with an active three-dimensional configuration. However, only occasionally do two wrong amino acids cancel out each other's faults. For example, double mutants containing A446 and A23, or A446 and A187, do not produce active enzyme. At this time, it does not seem wise to speculate on how the various amino acid residues are folded together in the three-dimensional configuration and why only some combinations are enzymatically active. This kind of analysis must await the establishment of the three-dimensional structure of tryptophan synthetase.

The Very Drastic Consequences of the Insertion or Deletion of Single Base Pairs[11,12]

Early on in the analysis of mutant proteins, it became clear that the vast majority of mutants being isolated did not yield the minimally altered proteins, bearing single amino acid replacements, that would arise through the change of one type of base pair into one of its three alternatives. Instead, most mutants represented changes that led to drastically altered gene products, often containing many fewer amino acids and with many of their amino acid sequences bearing no relationship to the wild-type polypeptide products. The nature of these mutants first became apparent through the proposal that such mutations usually represented either insertions or deletions of single nucleotide pairs. The drastic effect of these insertion or deletion events is a consequence of the fact that mRNA molecules are read in successive blocks of three nucleotides, called codons. AUG codons, which code for the methionine residues found at the amino terminal ends of newly synthesized polypeptide chains, are the signal for ribosomes to begin reading the mRNA molecule about to be translated into a protein. Since reading always begins at the appropriate AUG condon, the mRNA molecules are aligned on the ribosomes so that their messages are read in the correct **reading frame.**

If, however, a single base pair is inserted or deleted in a coding sequence, the triplets that designate amino acids become completely changed beginning at the site of insertion or deletion (Figure 8-18).

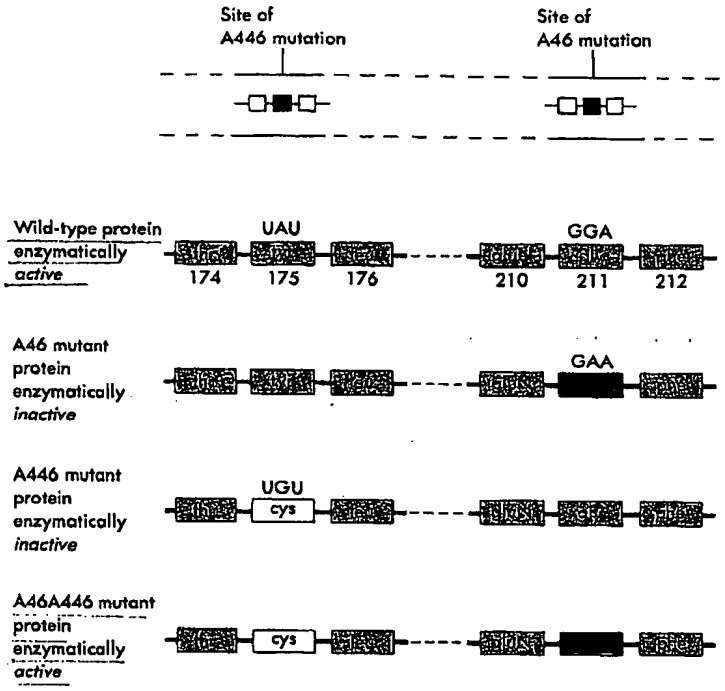


Figure 8-17
Reversal (suppression) of mutant phenotype by a second mutation at a second site in the same gene.

For example, if normally the gene sequence ATTAGACAC . . . is read as (ATT)(AGA)(CAC) . . . , then the insertion of a new nucleotide C in the fourth position of that sequence creates ATTCAGACAC, which is read as (ATT)(CAG)(ACA)(C . . .). These new triplets may code for entirely different amino acids. A similar consequence follows from a deletion. Moreover, the crossing of two deletion or two insertion mutants yields double mutants in which the reading frame is still misplaced.

Reversion of Insertion or Deletion Mutants

Active (or partially active) genes are regenerated by crossing over between an insertion and a nearby deletion. Such events restore the correct reading frame except in the short region between the mutations (see Figure 8-18). If the affected gene region is nonessential (e.g., the early section of the T4 *rIIB* gene), then the resulting protein product is fully functional. In other cases, the short segments of inappropriate amino acids are only mildly disadvantageous, and partial activity results. No activity, however, will usually be found if the inappropriate codons include any of the three that signify chain termination (UAA, UAG, or UGA). Their presence inevitably results in incomplete fragments of the wild-type polypeptide.

It is also sometimes possible to obtain functional genes by producing recombinants containing three closely spaced insertions or deletions (Figure 8-19). In contrast, recombinants containing four nearby insertions or deletions produce only nonfunctional polypeptides. These later experiments were performed in 1961, before the basic outlines of the genetic code were known. They in fact provided the first good evidence that the genetic code was likely to be read in groups of three as opposed to groups of two or four.

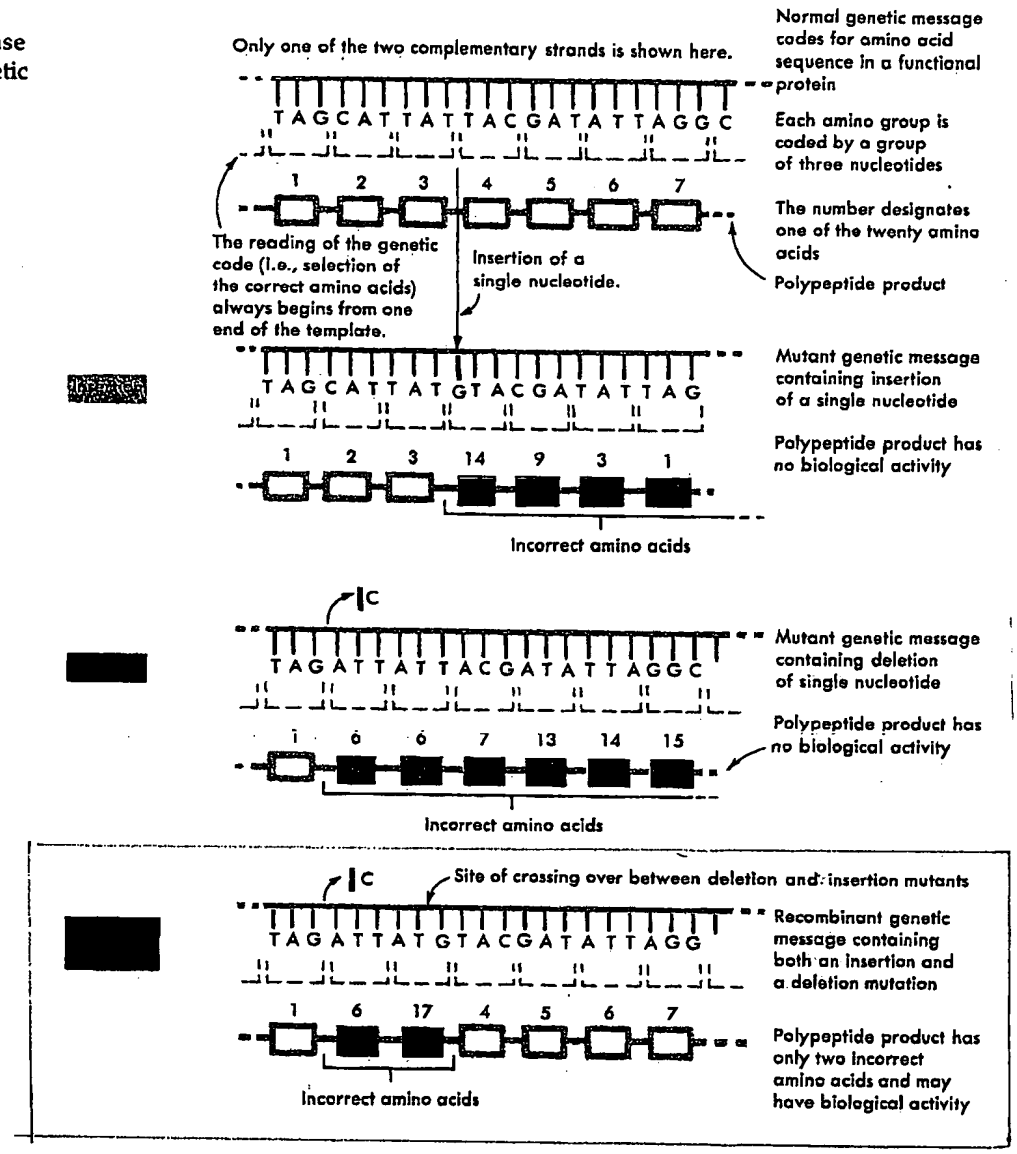
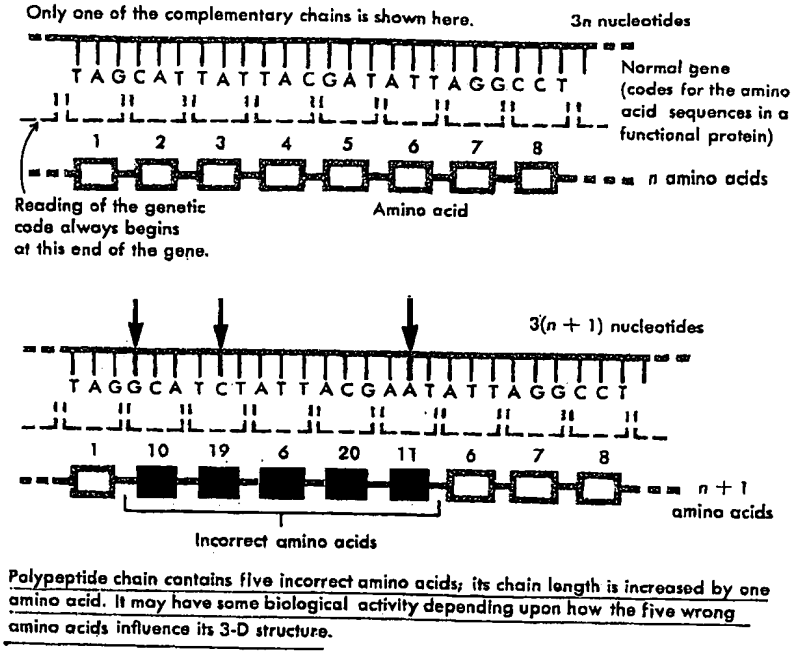


Figure 8-18
Mutations that add or remove a base shift the reading frame of the genetic message.

Cloned Genes Can Be Sequenced[13–17]

Virtually all the essential features of the genetic code were deduced by 1966 from the coding properties of either enzymatically or chemically synthesized mRNA molecules and from the accumulated knowledge of genetic fine structure that we have just detailed. No real genes were directly analyzed, however, since at that time there were no procedures either to sequence DNA or to isolate desired genes. But with the arrival of recombinant DNA and of powerful methods for DNA sequencing, the nature of genetic research has dramatically changed. No longer are genetic crosses the prime vehicle for probing genes. The quickest and most direct way to proceed is now the cloning and sequencing of relevant genetic material. As indicated in the previous chapter, it is now a relatively straightforward matter to isolat any *E. coli* gene that codes for a function that can be selected for by one of the many enrichment procedures.



Polypeptide chain contains five incorrect amino acids; its chain length is increased by one amino acid. It may have some biological activity depending upon how the five wrong amino acids influence its 3-D structure.

Figure 8-19
When three nucleotides are added close together, the genetic message is scrambled only over a short region. The same type of result is achieved by the deletion of three nearby nucleotides.

Already, a large number of *E. coli* genes have been completely or partially sequenced. In all cases, the codons found to specify given amino acids are those predicted by the genetic code (Figure 8-20). This agreement between prediction and result, though inherently very satisfying, surprised no one, since the experimental evidence used to deduce the genetic code was effectively unassailable (see Chapter 15). Also as predicted, the coding segments of virtually all mRNAs start with the AUG codon and always conclude with a chain-terminating codon (UAA, UAG, or UGA).

Untranslated Sequences at the Beginnings and Ends of mRNA Molecules[18–23]

When mRNA was first discovered, it seemed simplest to assume that the translation events would begin at one end of the molecule and then move along in steps of three nucleotides until the other end was reached. This was a very naive view, adopted before the discoveries that methionine initiates all polypeptide chains and that specific codons specify chain termination. Now we realize that untranslated sequences exist at both the 5′ end of the mRNA, near which translation begins, and at the 3′ end, near which translation stops (Figure 8-21). Hence, there must be internal signals in mRNA that mark the starting and stopping sites for translation. With the exception of a small purine-rich block of nucleotides that functions to position ribosomes at the correct AUG start codon, the untranslated regions probably play no role in translation and are of variable lengths, ranging from 20 to more than 100 nucleotides, depending on the particular mRNA species.

These seemingly unnecessary extra sequences only make sense